(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: 10124533 A
(43) Date of publication of application: 15.05.1998

(51) Int. Cl G06F 17/30
G06F 12/00

(21) Application number: 09121367
(22) Date of filing: 13.05.1997
(30) Priority: 13.05.1996 US 96 644599

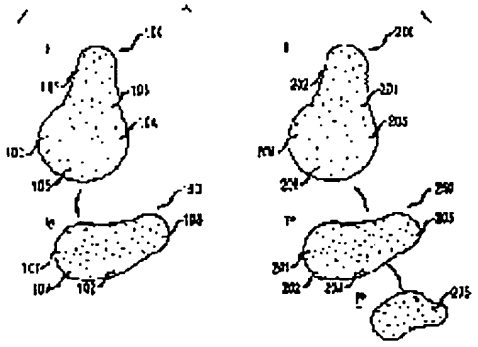
(71) Applicant: LUCENT TECHNOL INC
(72) Inventor: GANGULY SUMIT
GIBBONS PHILLIP B
MATIAS YOSSI
SILBERSCHATZ ABRAHAM

(54) METHOD FOR EVALUATING BIAS PREVENTION CONNECTION SIZE

(57) Abstract:

PROBLEM TO BE SOLVED: To obtain a method for evaluating a data base question sensitive to the possibility of the bias of data by applying an evaluating procedure to each defined group, and deciding the whole evaluation of a data base.

SOLUTION: A first data base R and a second data base T are respectively provided with plural data items 101-103 and 201-203 having different coupling attribute values. The evaluation is operated to the data items 101-103 and 201-203 having the specific coupling attribute densely occupying the both data bases R and T. Then, the evaluation for suppressing the influence of the dense data items 101-103 having the specific coupling attribute value in the first data base R is operated, and finally the evaluation for suppressing the influence of the dense data items 201-203 having the same coupling attribute value in the second data base T is operated. Then, the dense - dense and non-dense - any relation is defined by using the threshold value of each data base R and T.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平10－124533

(43)公開日　平成10年(1998)５月15日

(51)Int.Cl.[6]		識別記号	FI		
Ｇ０６Ｆ	17/30		Ｇ０６Ｆ	15/403	３３０Ｚ
	12/00	５１３		12/00	５１３Ｚ
				15/403	３４０Ｄ

審査請求　未請求　請求項の数13　ＯＬ　（全８頁）

(21)出願番号　特願平9－121367

(22)出願日　平成９年(1997)５月13日

(31)優先権主張番号　０８／６４４５９９

(32)優先日　1996年５月13日

(33)優先権主張国　米国（ＵＳ）

(71)出願人　596092698
ルーセント　テクノロジーズ　インコーポレーテッド
アメリカ合衆国．07974－0636　ニュージャーシィ，マレイ　ヒル，マウンテン　アヴェニュー　600

(72)発明者　サミット　ガングリー
インド国　462024　ボーパル，シャクティナガー，セクター－２　192

(74)代理人　弁理士　岡部　正夫　（外９名）

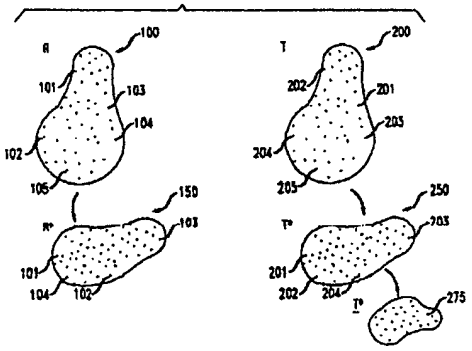
最終頁に続く

(54)【発明の名称】　偏り防止結合サイズ評価方法

(57)【要約】

【課題】　２つのデータベースＴ及びＲの質問サイズの評価方法を提供する。

【解決手段】　この方法は、データベースを稠密または疎であるとして類別するためにスレショールドを使用する。そこで、２つのデータベースに稠密－稠密手順が適用され、稠密－稠密評価（A_d）を作り出す。データベースＴからくるデータアイテムを抑制する疎－何れか手順が行なわれ、第１の疎－何れか評価（A_{o1}）を作り出す。次いで、データベースＲから稠密なデータアイテムを抑制することによって、第２の疎－何れか評価（A_{o2}）が作り出される。最後に、稠密－稠密評価、第１の疎－何れか評価及び第２の疎－何れか評価を結合することにより、質問サイズ評価が作り出される。



【特許請求の範囲】

【請求項１】　少なくとも２つのデータベースＲ及びＴの等結合を評価することによりデータを管理する方法であって、

Ｒの稠密なデータアイテムとＴの稠密なデータアイテムの質問サイズを評価することにより、稠密－稠密評価を作り出す工程と、

Ｒの稠密なデータアイテムを抑制する質問サイズを評価することにより、第１の疎－何れか評価を作り出す工程と、

Ｔの稠密なデータアイテムを抑制する質問サイズを評価することにより、第２の疎－何れか評価を作り出す工程と、

前記稠密－稠密評価と、前記第１の疎－何れか評価と、前記第２の疎－何れか評価を結合することにより、データベースＲ及びＴの等結合のサイズの評価を作り出す工程とからなる方法。

【請求項２】　請求項１記載の方法において、前記結合工程は、前記稠密－稠密評価と、前記第１の疎－何れか評価と、前記第２の疎－何れか評価を加算することにより行われる方法。

【請求項３】　請求項１記載の方法において、前記結合工程は、前記稠密－稠密評価と、前記第１の疎－何れか評価と、前記第２の疎－何れか評価を平均することにより行われる方法。

【請求項４】　請求項１記載の方法において、前記結合工程は、前記稠密－稠密評価と、前記第１の疎－何れか評価と、前記第２の疎－何れか評価のうちの１つを最大値として選択することにより行われる方法。

【請求項５】　少なくとも２つのデータベースＲ及びＴのデータベース質問サイズを評価することによりデータを管理する方法であって、

データベースＲにおけるある属性値を有する稠密なデータアイテムと前記属性値を有するＴの稠密なデータアイテムとの評価を行なうことにより、稠密－稠密評価を作り出す工程と、

前記データベースＲの稠密なデータアイテムを抑制する評価を行なうことにより、Ｒにおける疎－何れか評価を作り出す工程と、

Ｔの稠密なデータアイテムを抑制する評価を行なうことにより、Ｔにおける疎－何れか評価を作り出す工程と、

前記稠密－稠密評価と、前記Ｔにおける疎－何れか評価と、前記Ｒにおける疎－何れか評価を結合することにより、前記データベースＲ及び前記データベースＴのデータベース質問サイズを評価する工程とからなる方法。

【請求項６】　請求項５記載の方法において、前記稠密－稠密評価は、

前記データベースＲからデータアイテムをサンプリングすることにより、サンプルＲ＊を作り出し、

前記データベースＴからデータアイテムをサンプリングすることにより、サンプルＴ＊を作り出し、

前記サンプルＲ＊におけるある結合属性値（ｖ）を有する多数のデータアイテムを決定することにより、Ｒにおいて多数の前記結合属性値（ｖ）を作り出し、

前記サンプルＴ＊における前記結合属性値（ｖ）を有する多数のデータアイテムを決定することにより、Ｔにおいて多数の前記結合属性値（ｖ）を作り出し、

前記結合属性値（ｖ）の各々に関して、前記結合属性値（ｖ）のサブ結合のサイズの中間の稠密－稠密評価を決定し、

前記結合属性値（ｖ）の各々に関して中間の稠密－稠密評価を加算し、

前記サブ結合属性値（ｖ）の前記サイズの中間の稠密－稠密評価を見積もることにより行なわれる方法。

【請求項７】　請求項６記載の方法において、前記中間の稠密－稠密評価は、スレショールド値を確定し、Ｔにおける多数の前記結合属性値（ｖ）及びＲにおける多数の前記結合属性値（ｖ）を前記スレショールド値と比較した後に得られる方法。

【請求項８】　請求項７記載の方法において、前記中間の稠密－稠密評価は、Ｔ＊における多数の前記結合属性値（ｖ）及びＲ＊における多数の前記結合属性値（ｖ）が共に前記スレショールド以上であることを決定した後に得られる方法。

【請求項９】　請求項７記載の方法において、前記中間の稠密－稠密評価は、Ｔ＊における多数の前記結合属性値（ｖ）及びＲ＊における多数の前記結合属性値（ｖ）が共に前記スレショールドに等しいことを決定した後に得られる方法。

【請求項１０】　請求項５記載の方法において、前記第１の疎－何れか評価は、

前記データベースＴからデータアイテムをサンプリングすることにより、サンプルＴ＊を作り出し、

Ｒにおけるその数がスレショールド以上であるＴ＊のデータアイテムを抑制し、

前記結合属性値（ｖ）の各々について、Ｔにおいて疎である、前記結合属性値（ｖ）を有するデータアイテムにより中間の疎－何れか評価を計算し、

前記結合属性値（ｖ）の各々について中間の疎－何れか評価を加算し、

前記中間の疎－何れか評価を見積もることにより、ＲにおけるＴの疎－何れか評価を作り出すことによって行なわれる方法。

【請求項１１】　請求項１０記載の方法において、Ｔ＊のデータアイテムを抑制する前記工程は、稠密な結合属性値（ｖ）を有する、Ｒにおける多数のデータアイテムを決定することによって行なわれる方法。

【請求項１２】　請求項１０記載の方法において、Ｔ＊のデータアイテムを抑制する前記工程は、Ｒからランダムサンプルを＊をとり、Ｒ＊に表われる各結合属性値

（ｖ）に関して、Ｔ* から結合属性値（ｖ）を有する全てのデータアイテムを削除することにより行なわれる方法。

【請求項１３】　請求項５記載の方法において、前記第２の疎ー何れか評価は、

前記データベースＲからデータアイテムをサンプリングすることにより、サンプルＲ* を作り出し、

前記結合属性値（ｖ）を有する、Ｒにおける多数のデータアイテムを計算するコストを決定し、

Ｒ* の各データアイテムについて、前記結合属性値（ｖ）を有するＴのデータアイテムの数を決定し、

前記結合属性値（ｖ）の各々について、Ｒにおいて疎である、前記結合属性値（ｖ）を有するデータアイテムにより中間の疎ー何れか評価を計算し、

前記結合属性値（ｖ）の各々について中間の疎ー何れか評価を加算し、　前記中間の疎ー何れか評価を累積することにより、ＴにおけるＲの疎ー何れか評価を作り出すことによって行なわれる方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、データベース質問評価に関する。

【０００２】

【従来の技術及び発明が解決しようとする課題】コンピュータのまん延につれて、コンピュータデータベースも増加した。近頃のデータベースのサイズは非常に大容量になることがあり、データベースの中には数百乃至数十億のデータアイテムを保持しているものがある。これらのデータベースの内の１つのデータベースの質問では、あらゆるデータアイテムが質問に合う可能性があり、あらゆるデータアイテムを比較しなければならないことがある。したがって、これらのデータベースのサイズが増すにつれて、質問を実行するコストも増加している。

【０００３】データベースは１つ以上のテーブルからなり、各テーブルは数百乃至数十億のデータ記録を保持している。各データ記録は、情報が入っている１つ以上のフィールドを含む。これらのフィールド内の情報に基づいて、記録をいくつかのタイプのうちの１つのタイプとして類別することができる。例えば、テーブルは人間の記録を入れることができ、各記録は、人名を与えるフィールドと、好きなスポーツを与えるフィールドを有する。各記録は好きなスポーツで類別することができるので、タイプ“野球”、“フットボール”または“ホッケー”とすることができる。

【０００４】データベースのテーブルの質問に早く応答することができるのが望ましい。共通の質問の１つは等結合質問である。２つのテーブルＲ及びＴの等結合質問の結果は、１つの記録がＲからかつもう１つの記録がＴからのものであっていずれも同タイプのものである記録ペアの全てからなるテーブルとなる。例えば、テーブルＲが男性及び彼らが好きなスポーツの記録を含み、かつテーブルＴが女性及び彼女らが好きなスポーツの記録を含む場合は、テーブルＲ及びＴの等結合は、好きなスポーツが同じ男性と女性の記録ペアの全てを含むテーブルとなる。

【０００５】等結合の結果を計算するのは、コストがかかり過ぎることになる可能性がある。例えば、一方のテーブルがｎ個の記録を持ち、かつ他方のテーブルがｎ個の記録を持っている場合は、結果の計算はn^2の記録ペアの比較を必要とし得る。n^2の比較をそれぞれ実行すると、質問コストが増加する。したがって、等結合質問のコストを下げるのが好適である。

【０００６】大容量データベースにおける費用のかかる質問の実行のコストを減らす必要性の結果として、データベース評価の分野がポピュラーになってきた。データベース評価において、評価は、データベースにおける質問の可能な出力（質問の評価と呼ばれる）で作られる。したがって、質問の評価は、質問を行なう前に計算される。その結果、質問を続けてコストを負うべきかまたはこの特定の質問を取り消すべきかに関して決定することができる。データベース評価の問題点は、評価が正確かつ計算が能率的になるように、特定の質問の評価を計算することを伴うことである。

【０００７】このデータベース評価の問題点を解決する試みが以前に行われた。パラメータ法と呼ばれる手法は、データをとり、データベースにおいてこのデータと既知のデータモデルを比較するのを試みる。データモデルの作用と性質がデータベースのデータと同じであることが原理になっている。しかしながら、パラメータ法では、データが既知のデータモデルにどのくらい近似しているかについての仮定が行われる必要があり、近似（データの適合）は結果の正確さをはなはだしく変えてしまうことがある。他の種類の手法は、データベースにおけるデータアイテムのサンプル（小さな一組）をとり、これらのサンプルに基づいて評価を行なう。この種の手法はサンプリング法として知られている。サンプルは、通常、個々のデータベースの中からとられ、次いで合成されて質問評価を作り出す。サンプリング法は、他のデータベース評価法に勝る利点があることが証明された。パラメータ法と違って、サンプリング法は、データの適合について仮定がより少なく行われるべきことを要する。さらに、サンプリング法は、常に、統計的確実性と呼ばれる正確さの確実性を有する。たいていのサンプリング法の統計的確実性は、典型的に、９０％乃至９９％の範囲になるのを目指している。

【０００８】サンプリング法もパラメータ法もデータベース評価には好適ではない。しかしながら、サンプリング法は、データベースのデータアイテムが偏っている場合、パラメータ法より正確な結果を提供することができる。例えば、上記に説明したデータベースが同型でない

（同型でないデータベースとは、データベース内のデータアイテムが異なるタイプからなる同等でない混合になっているデータベースである）場合、パラメータ法は良好な性能を示すことができない。他のタイプより１つのタイプのデータアイテムがかなり多い（例えば２０％以上）場合、データベースは偏っているといわれる。偏ったデータベースは、パラメータ法やサンプリング法を実行する場合に問題があることがわかっている。

【０００９】サンプルが多くの偏ったデータアイテムを含んでいるかまたはあまり偏っていないデータアイテムを含んでいるかに依存して、偏ったデータの影響は、サンプリング法の結果に劇的に影響を与えることがある。サンプリング法を使用する場合、偏ったデータの影響を考慮することができない大量のサンプルがテーブルＴ及びテーブルＲからとられる。したがって、偏ったデータを考慮してサンプリング法を実行するのは不具合がある。

【００１０】

【課題を解決するための手段】本発明は、データの偏りの可能性に敏感な、データベース質問の評価方法を実行する。データベース質問の総合評価は、３つの個別評価を合わせて行われる。まず、評価は、稠密に両データベースＲ及びＴを占める、特定の結合属性を有するデータアイテムについて行われる。次に、第１のデータベースＲにある、特定の結合属性値を有する稠密なデータアイテムの影響を抑制する評価が行われ、最後に、第２のデータベースＴにある、同じ結合属性値を有する稠密なデータアイテムの影響を抑制する評価が行われる。

【００１１】本発明は、データベース質問の評価へ広範囲にわたるアプローチをとっている。この広範囲にわたるアプローチは、２つの異なるデータベースＲ及びＴのデータアイテムの関係を確立することによって作り出される。この関係は、サブ結合の収集からなる２つの部分に分かれたグラフとして知られている。各サブ結合は、この特定の結合属性値を有するデータアイテムの全ペアからなる。

【００１２】各データベースのランダムなサンプルが収集される（例えば、Ｒ* 及びＴ* は、それぞれデータベースＲ及びＴのランダムなサンプルを示す）。次いで、ランダムサンプルＲ* 及びＴ* におけるサブ結合は、特定の結合属性値のデータアイテム数がスレショールド値以上か以下かに基づいて、特定の結合属性値のデータアイテムの稠密な母集団または特定の結合属性値のデータアイテムの疎な母集団を持つように評価される。本発明の方法では、データアイテムは、その結合属性値が稠密ならば稠密になり、その結合属性を有するデータアイテムの母集団が疎ならば疎になることがわかる。スレショールド値は、データベースのデータアイテム数の平方根として定義される。サブ結合の各ペアのデータアイテムが共に稠密な場合は、サブ結合は稠密－稠密と呼ばれる。サブ結合の各ペアにおいて、一方のデータアイテムが疎なものとして類別され、他方のデータアイテムが稠密または疎なものとして類別されている場合は、サブ結合は、疎－何れか（例えば、稠密－疎、疎－稠密または疎－疎）として類別される。

【００１３】次いで、一連の３つの評価がデータベースで行われる。手順はまず、稠密－稠密評価（A_d）を決定するためにランダムサンプルに適用される。次いで、Ｒの結合属性値を有する稠密データアイテムを抑制する疎－何れか（A_{S1}）評価が行われ、次に、Ｔの結合属性値を有する稠密データアイテムを抑制する疎－何れか（A_{S2}）評価が行われる。最後に、稠密－稠密評価と、Ｒの稠密データアイテムを抑制する疎－何れか評価と、Ｔの稠密データアイテムを抑制する疎－何れか評価とを結合することによって、データベース質問評価（Ａ）が作られる。

【００１４】稠密－稠密評価と疎－何れか評価は、特定のデータベース質問に対して、他のランダムサンプル（例えばＴ* ）と適合する可能性のある一方のランダムサンプル（例えばＲ* ）における結合属性値を有するデータアイテムを評価することにより行われる。上述のように、特定の結合属性値を有するデータアイテムの適合はこの方法ではサブ結合と呼ばれる。したがって、各結合属性値に関して、その値と関連したサブ結合のサイズが評価される。次いで、サブ結合の評価の和が全ての結合属性値に関して加算され、これらの和（例えば、稠密－稠密評価、疎－何れか評価）は組み合わさってデータベース質問評価を作り出す。

【００１５】本発明の目的、利点及び新規な特徴は、添付図面に関して読まれる以下の詳細な説明からより十分に明らかになるだろう。

【００１６】

【発明の実施の形態】本発明では、等結合データベース質問サイズの評価方法が実行される。まず、質問基準が定義される。各データベースからのデータアイテムは質問基準にしたがってグループ分けされる。次いで、“稠密－稠密”手順及び“疎－何れか”として知られる評価手順が、定義されたグループの各々に適用され、データベースの全体評価を決定する。

【００１７】詳細には、この方法は、２つのデータベースの稠密－稠密評価（A_d）を行なうことによってデータベースＲ及びＴの等結合評価（Ａ）を行なう。次いで、Ｒの稠密なデータアイテムを抑制する疎－何れか手順が行われ、第１の疎－何れか評価（A_{S1}）を作り出す。次いで、Ｔの稠密なデータアイテムを抑制する疎－何れか手順が行われ、第２の疎－何れか評価（A_{S2}）を作り出す。最後に、A_d，A_{S1}及びA_{S2}を組み合わせる（加算する）ことによって、２つのデータベースの等結合評価（Ａ）が計算される。

【００１８】本発明の方法を示す手段として、図１は、

説明を容易にするために用いることができる概念的モデルを表わしている。図１において、１００で示された第１のデータベース（R）及び２００で示された第２のデータベース（T）が示されている。第１のデータベースR及び第２のデータベースTは共に、異なる結合属性値（例えば、それぞれ１０１，１０２，１０３及び２０１，２０２，２０３）を有する多数のデータアイテムを有している。また、２つのデータベースサンプルが示されている。第１のランダムサンプル（R*）は１５０で示され、第２のランダムサンプル（T*）は２５０で示されている。両ランダムサンプル１５０及び２５０は、原データベースR及びTからのデータアイテムのランダムサンプリングを含んでいるだろう。したがって、１６０で示された第１のランダムサンプル（R*）はデータアイテム１０１，１０２及び１０３を含み、２５０で示された第２のランダムサンプル（T*）はデータアイテム２０１，２０２及び２０３を含む。

【００１９】図１に表わされた概念的モデルの直観的な理解を発展させるために、データベースRのデータアイテムは、好きなスポーツを持っている男性の母集団を表わすと仮定する（好きなスポーツは結合属性値となるだろう）。したがって、データアイテム１０１，１０２及び１０３は各々、特定のスポーツが好きな特定の男性を表わす。次に、データベースTは、好きなスポーツを持っている女性の母集団を表わすと仮定する。したがって、各データアイテム２０１，２０２及び２０３は、特定のスポーツが好きな女性の男性を表わす。さらに、簡単にするために、安静または女性の各々が好きなスポーツ（結合属性値）は、ベースボール、フットボールまたはホッケーのどれかであると仮定する。したがって、データベースRは、好きなスポーツがベースボール、フットボールまたはホッケーである男性の母集団を表わし、データベースTは、好きなスポーツがベースボール、フットボールまたはホッケーである女性の母集団を表わす。

【００２０】データベースR及びTの可能な質問（等結合）は、（１）ベースボールが好きな男性と女性、（２）フットボールが好きな男性と女性、（３）ホッケーが好きな男性と女性に適合することができる。このタイプの質問は、評価技術なしに行われる場合は３ステップで達成することができる。まず、データベースRは、好きなスポーツがベースボールである全男性についてサーチされるだろう。次に、データベースTは、好きなスポーツがベースボールである全女性についてサーチされるだろう。最後に、好きなスポーツがベースボール（フットボール及びホッケー）である男性と女性の好一対が作られる。データベースR及びTが共にnデータアイテムを含んでいる場合、好きなスポーツが同じである男生と女性を組み合わせるのはn^2の作業になる。

【００２１】本発明の方法論では、ランダムサンプルR*及びT*は、それぞれデータベース（R）及び（T）から取られる。次いで、サンプルは結合属性値で類別される。結合属性値でデータアイテムを類別するために、好きなスポーツは各々、それと関連する番号（例えば、ベースボール“１”、フットボール“２”及びホッケー“３”）を持っていると仮定する。そこで、特定のスポーツが好きな男性と女性の好一対は、図１の２つの部分に分かれたグラフで表わすことができる。図１Bにおいて、図１Aのランダムサンプル R* 及び T* からの各データアイテムは、その上に好きなスポーツの番号を伴って示されている。例えば、データアイテム１０１は、（データアイテム１０１上の数字が１なので）ベースボールが好きな、ランダムサンプル R* からの男性である。図１Bのデータアイテム２０１はランダムサンプル T* からのデータアイテムである。したがって、図１Bのデータアイテム２０１は、（その上に２があるので）フットボールが好きな女性である。

【００２２】そこで、特定のスポーツが好きな男性と女性をペアにして、２つに別れたグラフとして表わすことができ、このグラフには、各男性及び女性を表わすノードと、男性と女性が好きなスポーツが同じ場合の、男性を表わすノードと女性を表わすノード間の線（データアイテム間に引かれた線）とがある。図１Bにおいて、データアイテム１０１，１０２及び１０３は、（それぞれ、１，３及び２で表わされるスポーツが好きな）３人の男性の記録を表わし、データアイテム２０１，２０２及び２０３は、（それぞれ、２，１及び２で表わされるスポーツが好きな）３人の女性の記録を表わす。３０２，３０４，３０６で示された線は、１０３で示された男性（ベースボールが好きな男性）を（ベースボールが好きな）３人の女性とペアにする。

【００２３】一般に、各結合属性値（例えば、ベースボール、フットボール、ホッケー）について、この属性値（例えば１，２，３）を有する男性の各々を同じ属性値（例えば１，２，３）を有する女性とペアにする線がある。ある特定の結合属性値のデータアイテムと線を１組にして、サブ結合として示されている。正確には、各結合属性値（例えば１で示されたベースボール、２で示されたフットボール及び３で示されたホッケー）に対して１つのサブ結合がある。例えば、ベースボールが好きな男性と女性のサブ結合は、データアイテム１０３，１０４と、線３０２，３０３，３０４，３０５，３０６，３０７と、データアイテム２０１，２０３及び２０４とからなるだろう。

【００２４】個々のデータベースのスレショールド値を用いて、本発明は、稠密－稠密及び疎－何れか（すなわち、稠密－疎、疎－稠密及び疎－疎）のような関係を定義する。スレショールド値は、ランダムサンプルを稠密なデータアイテムの母集団の表現にすることが可能などんな値でも良いことがわかる。この関係は次のように定義

される。

【００２５】稠密－稠密－データベース（R）におけるある結合属性値を有するデータアイテムの数とデータベース（T）における同じ結合属性値を有するデータアイテムの数が各々、平方根（n）より大きいかまたは等しい関係。ここで、nはデータベースのデータアイテム数である。

【００２６】稠密－疎－データベース（R）におけるある結合属性値を有するデータアイテムの数が平方根（n）より大きいかまたは等しく、かつデータベース（T）におけるデータアイテムの数が平方根（n）より小さい関係。ここで、nは各データベースのデータアイテム数である。

【００２７】疎－稠密－データベース（R）における特定の結合属性値を有するデータアイテムの数が平方根（n）より小さく、かつデータベース（T）におけるデータアイテムの数が平方根（n）より大きいかまたは等しい関係。ここで、nは各データベースのデータアイテム数である。

【００２８】疎－疎－データベース（R）及びデータベース（T）における特定の結合属性値を有するデータアイテムの数が平方根（n）より小さい関係。ここで、nは各データベースのデータアイテム数である。

【００２９】本発明の方法論では、ある結合属性値vを有する１組のデータアイテムについて、“$mult_T$（v）”は、その結合属性値を有するTにおけるデータアイテムの数と定義される。Rにおいて、“$mult_R$（v）”がnの平方根より大きいかまたは等しければ、結合属性値は稠密であると定義され、また、“$mult_R$（v）”がnの平方根より小さければ、結合属性値が疎であると定義される。さらに、サブ結合（v）は、vがR及びTの両方で稠密ならば、稠密－稠密サブ結合となる。サブ結合（v）は、vがR及びTの両方で疎ならば、疎－疎サブ結合となる。

【００３０】本発明の方法論では、nは各関係におけるデータアイテム数であり、m_1は稠密－稠密手順のサンプルサイズであり、M_2は疎－何れか手順のサンプルサイズであり、δは稠密－稠密手順で使用されるスレショールド値である。上記に定義された変数を仮定すれば、各評価は次のように定義することができる。

【００３１】$A_d := f_d\ (n, m_1, \delta)$

$A_{s1} := f_s\ (R, T, n, m_2)$

$A_{s2} := f_s\ (T, R, n, m_2)$

$A := A_d + A_{s1} + A_{s2}$

【００３２】ここで、f_d（x）及びf_s（x）はxの関数であり、Aは総合データベース評価であり、A_dは稠密－稠密データベース評価であり、A_{s1}は、Rの稠密なデータアイテムを抑制する疎－何れかデータベース評価であり、A_{s2}はTの稠密なデータアイテムを抑制する疎－何れか評価である。$A < n\log n$ならば、この方法は、質問サイズの上限である健全な限界（S）：$=n\log n$も提供する。開示された実施例では、$m1 =$（平方根（n）$+\log n$）$*\log n$，m_2＝平方根（n）$+\log n$、$\delta = \log n$となる。

【００３３】本発明では、各データベースのランダムサンプルがとられる。２つのデータベースからとられたランダムサンプルはほとんど$m_1 + m_2$であり、m_1は、稠密－稠密評価を行なう場合に（個別的に）R及びTからサンプリングされたデータアイテム数であり、m_2は、疎－何れか評価を行なう場合に（個別的に）R及びTからサンプリングされたデータアイテム数である。サンプルがとられると、稠密－稠密評価（A_d）が引き出され、Rの稠密なデータアイテムを抑制する疎－何れか評価（A_{s1}）が計算され、Tの稠密なデータアイテムを抑制する疎－何れか評価（A_{s2}）が計算される。最後に、A_d，A_{s1}及びA_{s2}を結合することにより、質問サイズの評価を行なうことができる。

【００３４】稠密－稠密

この方法論、稠密－稠密手順の第１のステップは、以下のステップを含む。

１．データベースR及びTからランダムサンプルR* 及びT* がとられる。ランダムサンプルR* 及びT* は各々サイズm_1からなる。

２．V* はR* 及びT* の両方におけるデータアイテムの１組の結合属性値とする。

３．各値v⊂V* について、$mult_{R*}$（v）を決定する。

４．各値v∈V* について、$mult_{T*}$（v）を決定する。

５．各値v∈V* について、$mult_{R*}$（v）≧δかつ$mult_{T*}$（v）≧δならば、A'：$=$A'＋$mult_{R*}$（v）$*mult_{T*}$（v）となる。ここで、A'は中間の稠密－稠密評価（A'は初期にゼロに設定される）であり、“*”は乗算を示すために使用される記号である。

６．稠密－稠密評価Ad：$=$（n／m_1）^{2} $*$A'となる。

【００３５】疎－何れか

疎－何れか評価を作り出す方法は以下の手順からなる。

１．データベースTからランダムサンプルT* がとられる。ランダムサンプルT* はサイズm_2からなる。

２．$mult_{R*}$（v）が稠密vについて多数の計算を必要とするならば、稠密vを抑制するために蓋然論的消去を使用する。

２ａ．データベースRからランダムサンプルR* がとられる。ランダムサンプルR* はサイズm_2からなる。

２ｂ．R* に表われる各結合属性値について、T* から結合属性値vを有する全てのデータアイテムを削除する。

３．Rにおいて疎である残りの結合属性値に基づいて中

間の疎ー何れか評価を計算する。
T* における各データアイテムyについて、
3a．yと結合する、すなわちyと同じ結合属性値を有する、Rのデータアイテムの番号xを決定する。
3b．$x < n/m_2$ ならば、A'：＝A'＋xとなる。ここで、A'は中間の疎ー何れか評価である（A'は初期にゼロに設定される）。
4．A_s ：＝nA'／m_2 となる。ここで、A_s は疎ー何れか評価である。

【0036】本発明の方法論では、疎ー何れか評価は2度行なわれる。疎ー何れか手順が2回行なわれ、1回目に使用されなかったデータベースの稠密なデータアイテムが抑制される。例えば、データベースRの稠密なデータアイテムが抑制されたならば、1回目に疎ー何れか手順が行なわれ、2回目にデータベースTの稠密なデータアイテムが抑制されるだろう。1回目に疎ー何れか評価が行なわれると、その評価はA_{s1}で示され、2回目に（RとTの役割が反対にされて）疎ー何れか評価が行なわれると、その評価はA_{s2}で示される。稠密ー稠密及び疎ー何れかの評価の全てが計算されると、等結合評価は、$A = A_d + A_{s1} + A_{s2}$で示される。A＜（n）log（n）ならば、健全な限界S＝（n）log（n）の出力を計算することができる。ここで、Sは健全な限界である。すなわち、ハットは評価の統計的確信を維持する。さらに、等結合評価を計算するために、稠密ー稠密評価、第1の疎ー何れか評価及び第2の疎ー何れか評価を追加するべく、等結合評価は、3つの評価を平均するかまたは3つの評価の最大をとることにより近似することもできることがわかる。

【0037】さらに、本発明の方法論は、どんな数のデータベースにも適用することができる。例えば、3つのデータベースA，B及びCが含まれている場合は、この方法論は、まずデータベースA及びBに適用して中間の評価を作り出し、次いで、この方法論をCに用いて中間評価を結合することができる。

【0038】また、本発明は、データベース質問の正確で費用のかからない評価を作り出すために適用することができる。この方法は、質問最適化装置に、または並列もしくは分散データベースの多数のプロセッサにおける仕事量のバランスを取るのに必要なリソース割当ての決定時に有効である。さらに、より広範囲の規模に、開示されたこの方法を、会計監査や統計的研究等の大容量データベースアプリケーションにも適用することができる。

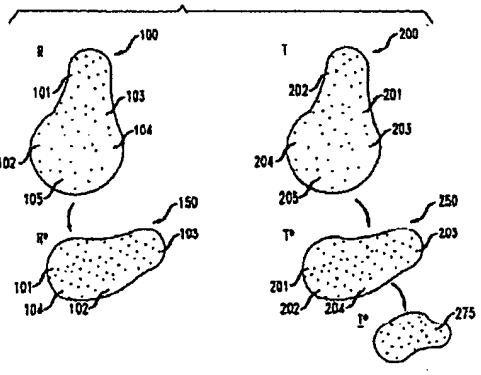
【0039】本発明のいくつかの実施例が開示されて説明されているが、本発明の精神または従属請求項の範囲から逸脱することなく種々の変更を行なうことができることがわかる。

【図面の簡単な説明】

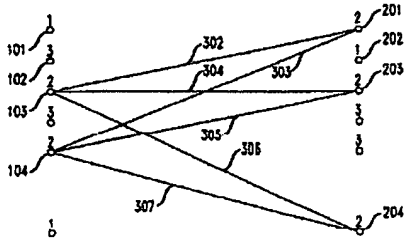
【図1A】母集団R及びTとランダムサンプルR* 及びT* の概念図を示す。

【図1B】ランダムサンプルR* 及びT* の選択されたデータアイテムと、結合属性値2のサブ結合の2つの部分に分かれたグラフを示す。

【図1A】



【図1B】



フロントページの続き

(72)発明者　フィリップ　ビー．ギボンズ
アメリカ合衆国　07090　ニュージャーシィ，ウエストフィールド，エンブリー　コート　201

(72)発明者　ヨッシ　マティアス
アメリカ合衆国　20854　メリーランド，ポタマック，ロザリンダ　ドライヴ　11815

(72)発明者　アブラハム　シルバーシャッツ
アメリカ合衆国　07901　ニュージャーシィ，サミット，ニューイングランド　アヴェニュー　67ユー